

AD-A241 634



Naval Research Laboratory

Washington, DC 20375-5000



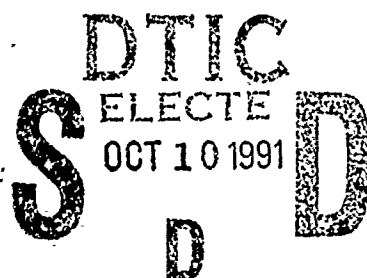
NRL Report 9340

Unlimited Vocabulary Synthesis Using Line Spectrum Pairs

STEPHANIE S. EVERETT

*Human-Computer Interaction Laboratory
Information Technology Division*

September 30, 1991



91-12927



Approved for public release; distribution unlimited.

91 10 9 047

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.</small>				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE September 30, 1991	3. REPORT TYPE AND DATES COVERED Final 10-86 thru 9-90		
4. TITLE AND SUBTITLE Unlimited Vocabulary Speech Synthesis Using Line Spectrum Pairs		5. FUNDING NUMBERS PE - 61153N WU - DN2573		
6. AUTHOR(S) Everett, Stephanie S.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Research Laboratory Washington, DC 20375-5000		8. PERFORMING ORGANIZATION REPORT NUMBER NRL Report 9340		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Space and Naval Warfare Systems Command Washington, DC 20363-5100		10. SPONSORING/MONITORING AGENCY REPORT NUMBER		
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.		12b. DISTRIBUTION CODE		
13. ABSTRACT (Maximum 200 words) <p>This report describes a new speech synthesis system based on the line spectrum pair (LSP) representation of the speech signal. The system includes a library of approximately 250 stored speech segments, represented by LSP parameters, that are concatenated to produce an unlimited output vocabulary. All components of the system are discussed, and particular emphasis is placed on the methods used to adjust the pitch, duration, and amplitude of concatenated segments to generate natural-sounding speech. A new method of computing fundamental frequency contours for sentential intonation patterns is presented. Test results show good segmental intelligibility for the synthesized speech: a score of 86.3 was obtained on the Diagnostic Rhyme Test (DRT); a score of 86.7 was obtained on the Modified Rhyme Test (MRT).</p>				
14. SUBJECT TERMS Speech synthesis LSP Synthetic speech		15. NUMBER OF PAGES 18		
		16. PRICE CODE		
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

CONTENTS

INTRODUCTION	1
SYSTEM DESCRIPTION	2
Segment Library	2
Linguistic Rules	6
Timing Rules	7
Intonation Contours	8
Amplitude Contours	10
Synthesis Algorithm	10
SYSTEM PERFORMANCE EVALUATION	10
CONCLUSIONS	11
ACKNOWLEDGMENTS	13
REFERENCES	13

Accession For	
NTIS CRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	



UNLIMITED VOCABULARY SPEECH SYNTHESIS USING LINE SPECTRUM PAIRS

INTRODUCTION

As computers continue to become more widely used and more powerful, the importance of efficient and easy-to-use human-computer interfaces becomes more apparent. Because speech is the most natural form of communication for humans, developers are now beginning to include speech input and output as part of the interface. Speech output can take two basic forms: the playback of digitally stored "canned" messages for applications such as telephone numbers, bank balances, and repeated phrases, and the generation of an unlimited vocabulary of synthesized utterances for more complex applications such as information retrieval, computer-assisted training, and aids for the handicapped.

The synthesis of an unlimited vocabulary from phonemic or orthographic representations has been studied for over 30 years (see Ref. 1 for an excellent review). Historically the most popular approach to this problem has been formant-based synthesis by rule. In this approach a set of rules generates a description of the time-varying patterns of spectral resonances (formants) for each sound in terms of the resonant frequency, bandwidth, and amplitude. The output speech is generated by passing the parameters through a set of bandpass filters. These systems are capable of generating relatively high-quality speech, but they require large and complex sets of rules. Because the rules interact, one of the main problems with these systems is the difficulty of writing good rules. Another problem is the difficulty of generating multiple voices without using multiple sets of rules.

Another approach to unlimited vocabulary synthesis is the concatenation of small segments excised from natural speech. Synthesis by concatenation requires far fewer rules than synthesis by rule, because the sounds are simply retrieved as needed. The segments are taken from natural speech and usually are stored as sets of linear predictive coding (LPC) prediction coefficients to facilitate adjustments to pitch, duration, and amplitude. Unfortunately it is difficult to make fine adjustments to the speech spectrum in an LPC-based system. Because the parameters are not in the frequency domain, a change in one parameter affects the shape of the entire spectrum at that point in time. This can cause distortion at segment boundaries, especially if trajectories must be significantly smoothed or adjusted.

The objective of this investigation is to determine the feasibility of using line spectrum pair (LSP) parameters instead of LPC parameters in a synthesis-by-concatenation system. The primary advantage of LSPs is that they are in the frequency domain. A change in one parameter affects the spectrum only in that frequency region, so it is easy to make fine adjustments to the speech spectrum. With this type of system, multiple voices can be synthesized easily by having multiple sets of stored segments.

LSP parameters are derived from LPC prediction coefficients through the decomposition of the impulse response of the LPC analysis filter into a pair of even and odd functions, each having roots along the unit circle of the complex z plane [2-4]. LSP parameters are naturally ordered and

are continuous, even across unvoiced sounds. Figure 1 shows the closely spaced parameter trajectories that correspond to peaks in the speech spectral envelope (formants) and the widely spaced trajectories that correspond to valleys.

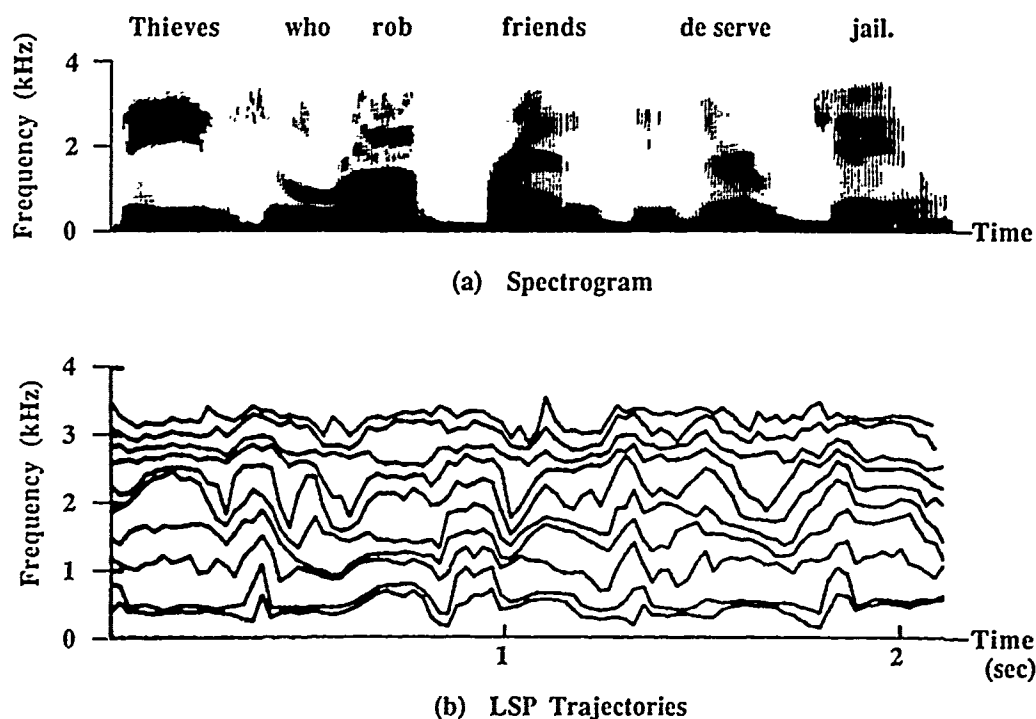


Fig. 1 — Typical LSP trajectories with a spectrogram of the original speech showing the formant trajectories for comparison. Closely spaced LSPs correspond to spectral peaks, or formants; widely spaced LSPs correspond to valleys.

SYSTEM DESCRIPTION

This synthesizer contains a library of segments that have been excised from natural speech and stored as sets of LSP parameters. For each utterance the specified segments are retrieved from the library and concatenated. By using stress levels entered by the user, the timing and duration of each segment are adjusted by a set of context-sensitive rules, and intonation and amplitude contours are computed for the entire utterance. The output speech is then generated by an enhanced LSP/LPC synthesis algorithm. Figure 2 outlines this process.

At the present time, this synthesizer is implemented in FORTRAN on a MicroVax 3600, using a DSC-200 A-D/D-A converter. It comprises approximately 1000 lines of code and requires roughly 200 Kbytes of memory to store the segment library. Computation requires approximately 2.8 s for each second of speech, plus the time needed to load the D-A buffers and generate the output (usually 4 to 5 s). Because this is a developmental system, no efforts have been made to optimize execution speed.

Segment Library

The library contains approximately 250 segments taken from natural speech. A list of words was read by one male speaker and digitized at an 8 kHz sampling rate using a 4 kHz low-pass anti-aliasing filter. LSP parameters for each word were computed and stored in separate files; an interactive display and editing program was used to locate and excise the desired segments. Other LPC-based synthesis-by-concatenation systems use diphone segments that extend from the midpoint of one phoneme to the midpoint of the next [5, 6]. In this system we chose to use a phoneme-based approach because of the ease with which LSP trajectories and spectra can be

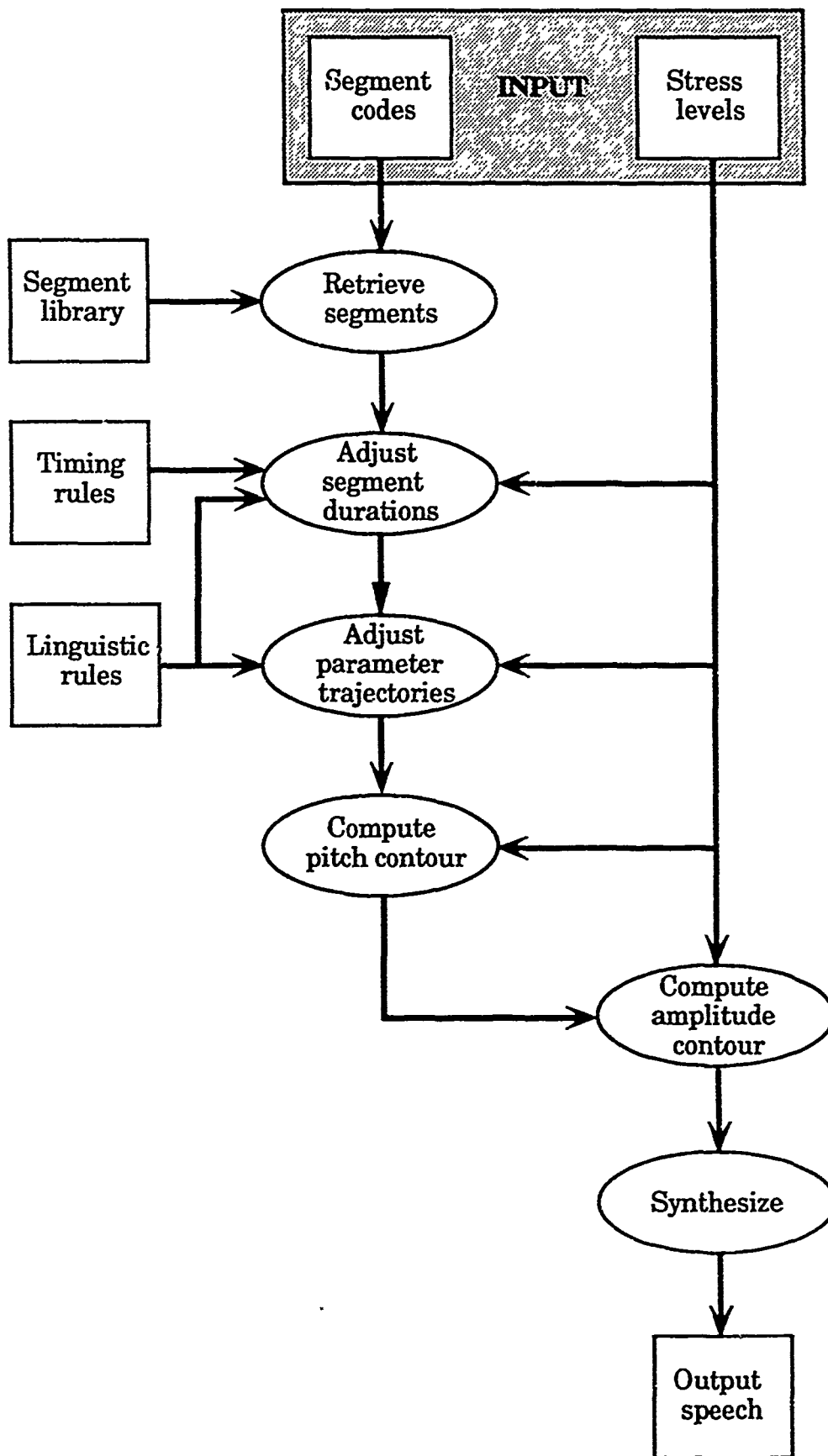


Fig. 2 — Overview of the LSP-based synthesizer described in this report

adjusted. (LSPs would also be useful in a diphone concatenation system.) Segment boundaries were determined by visually inspecting the LSP trajectories and by listening. Clear, normal pronunciations of each phoneme were used. Repeated listening and extensive testing of phoneme combinations were performed to verify that the segments chosen for the library were correctly delimited. Manual adjustments to parameter trajectories were made as necessary to minimize coarticulation and to remove irregularities.

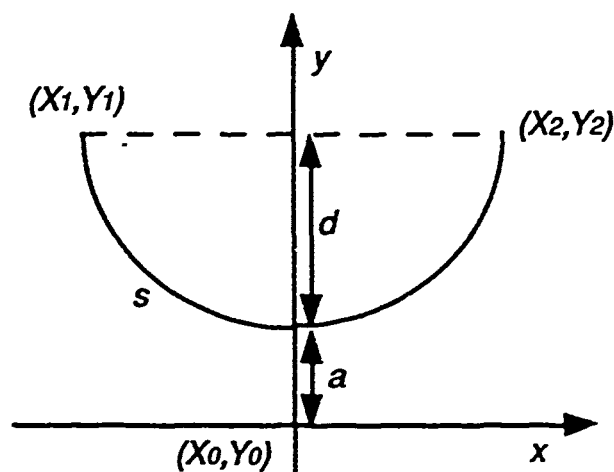


Fig. 3 — A catenary curve

Once the desired segments have been extracted from the original speech, the amplitude contours are normalized. For each segment, the maximum amplitude is located, and the amplitude for each frame of that segment is converted to a percentage of the local maximum value. In this way the original amplitude contour of each segment is maintained, and the relative amplitudes of the segments can be easily adjusted.

The segments used in this system range in size from subphonemic units, such as the glottal stop at the beginning of utterance-initial vowels, or the *W*-like offglide in *how*, to groups of two or three phonemes, such as word-initial consonant clusters or certain vowel-liquid sequences. The distribution of segments used was determined from listening tests. To generate natural-sounding speech it is necessary to store allophonic variations of most sounds (see the segment inventory in Table 1). Because some allophones occur in only one environment (e.g., consonants in clusters; vowels before *R* or *L*), the allophone and its environment are stored as a unit, thus reducing the segment inventory and simplifying the retrieval and concatenation processes. The use of segments of varying lengths reduces the need for complex combinatory rules and produces more intelligible speech.

Table 1(a) lists the consonant segments in the LSP library. Most consonants have two variants for use in syllable-initial position: one used before front vowels (*EE, IH, EY, EH, AE*) and one used before all other vowels. This is because the first and second formants of front vowels are widely separated, whereas for back vowels they are closer together. Using two allophones produces more natural coarticulation and generates more intelligible speech. The phoneme *H* is more strongly affected by the following vowel than are other consonants, so five variants are needed instead of two. The segment *H1* is used before high back vowels (*UW* and *UH*, as in *HOOK*); *H2* is used before non-high back vowels (*OW, OR, OY*); *H3* is used before low vowels (*AA, AR, AU, AY, AO, UX*). *Hf1* is used before high front vowels (*EE, IH*); *Hf2* is used before non-high front vowels (*EY, EH, ER, AE*) and schwa (*EX*). *Y1* is used before high vowels, both front and back, and *Y2* is used before all non-high vowels. Satisfactory syllable-final consonant clusters can be generated by concatenating individual phonemes. However, for syllable-initial clusters, better results are obtained when the clusters are stored as units.

Table 1(a) — Consonant Segments Contained in the LSP Synthesizer Library

IPA Symbol	Syllable-Initial (standard form)	Syllable-Initial (front vowels only)	Initial Clusters	Syllable Final	Syllabic Consonants
b	B	Bf	BR, BL	B	
p	P	Pf	PR, PL	P	
d	D	Df	DR	D	
ð	TD				
t	T	Tf	TR	T	
g	G	Gf	GR, GL	G	
k	K	Kf	KR, KL, KW	K	
θ	DH			DH	
θ	TH			TH	
v	V			V	
f	F		FR, FL, FLf	F	
z	Z			Z	
s	S		SL, ST, STR, SP, SPR, SW, SWf	S	
ʃ				ZH	
ʃ	SH			SH	
dʒ	J	Jf		J	
tʃ	CH			CH	
m	M	Mf		M	MM
n	N	Nf		N	NN
ŋ				NG	
r	R				
l	L	Lf			.LL
w	W	Wf		W	
h	H1, H2, H3	Hf1, Hf2			
j	Y1, Y2	Y1, Y2		Y	

Table (1b) lists the vowel segments in the library. The formant structure of the following consonant affects lax vowels (*IH*, *EH*, *UX*) more strongly than other vowels, so additional allophones of those sounds are included. Segments ending in 1 are used before front (labial) consonants (*P*, *B*, *F*, *V*, *M*); those ending in 2 are used before mid (dental and alveolar) consonants (*T*, *D*, *TH*, *DH*, *S*, *Z*, *CH*, *J*, *N*); those ending in 3 are used before back (palatal and velar) consonants (*K*, *G*, *SH*, *ZH*). The nasalized forms of all vowels are used before *M*, *N*, and *NG*, unless there is a special segment for that vowel for use before *NG* (*ING*, *ANG*). Because these allophones of *IH* and *AE* occur only before *NG*, the nasal is included in the segment. Likewise, all the *L*- and *R*-colored vowel segments include the liquid. The onset segments are very short, composed of the release of the glottal stop and only one or two frames of the vowel. These segments are inserted preceding the vowel segment for utterance-initial vowels. The segments *IHY*, *EHY*, *AEY* and *UXY* contain a *Y*-like off-glide and are used only before *G*. The sounds *ER*, *AR*, and *OR* are treated as units rather than as vowel+*R* sequences because they follow the same distribution pattern as the regular vowels, and because this allows complex segments such as *ORL* to be stored as a single segment.

Each segment in the library contains a header that includes certain invariant information about the segment, including the length of the file, phonemic classification, and the factors used in adjusting segmental timing and amplitude. Specifically, the headers comprise the following information:

- Segment length — the number of sets of LSP parameters (frames) contained in the file.

- Beginning and end of subsegment — the first and last frames of a shorter version of the segment that is used for vowels preceding unvoiced consonants and for consonants in clusters.
- Phonemic class — whether the segment is a vowel, liquid, nasal, fricative, plosive, or pause. Phonemic class is indicated for both the beginning and the end of the segment to accommodate mixed class segments (e.g., the segment *TR-* begins with a plosive but ends with a liquid).
- Segment voicing — the presence or absence of periodic excitation. For segments containing more than one phoneme, the voicing characteristic of the first phoneme is stored.
- Maximum compression allowed (see the discussion of timing rules).
- Amplitude adjustment factor for the segment as a whole (see the discussion on amplitude contours).
- Point and height of articulation.

Table 1(b) — Vowel Segments Contained in the LSP Synthesizer Library

IPA	Segment	Nasal	L-color	R-color	Onsets	Other
i	EE	EEN	EEL	EER	#EE	
I	IH1	IHN	IHL		#IH	IHY, ING
I	IH2					
I	IH3					
e	EY	EYN	EYL	EYR	#EY	
ɛ	EH	EHN	EHL		#EH	EHY
ɛ	EH2					
ə	EX	EXN			#EH	
ə	ER	ERN	ERL		#EH	
æ	AE	AEN	AEL	AER	#AE	AEY, ANG
a	AA	AAN	AAL		#AA	
ɹ	AR	ARN	ARL		#AA	
au	AU	AUN	AUL	AUR	#AA	
ai	AY	AYN	AYL	AYR	#AA	
ɔ	AO	AON	AOL		#AO	
ʌ	UX1	UXN	UXL		#UX	UXY
ʌ	UX2					
ʌ	UX3					
o	OW	OWN	OWL		#OW	
ɔr	OR	ORN	ORL		#OW	
oi	OY	OYN	OYL		#OW	
ʊ	UH	UHN	UHL		#UW	
u	UW	UWN	UWL	UWR	#UW	
iu	YU					

Linguistic Rules

Once the specified segments are retrieved from the library they are scanned by a set of context-sensitive linguistic rules. These rules are very similar in effect to rules or classes of rules found in formant-based synthesis-by-rule systems. They are ordered, and more than one rule may apply to a given segment. Segments are scanned in a single pass from left to right. Information about the current segment and the phonemic attributes of the following segment are obtained from the segment headers to determine the applicability of each rule to each segment.

In English, one of the primary cues to the voicing of a final consonant is the length of the preceding vowel [8, 9]. In other synthesis-by-concatenation systems, vowel length is adjusted by extending the stored segment through the repetition of data [10] or time warping [6]. In this

system vowel length adjusted by shortening the segment to use a subset of the data stored for each vowel. To determine if vowel length adjustment is necessary, the first linguistic rule checks to see if the vowel segment is followed immediately by an unvoiced consonant in the same word. If it is, the vowel segment is shortened to use the subsegment specified in the header. For example, in the word *bat* there are three segments: *Bf*- (*B* before a front vowel), *AE*, and *-T*. As stored in the library, *AE* contains 12 sets (frames) of data (195 ms). This is an appropriate length for use before a voiced consonant such as the *D* in *bad*, but it is too long for use before the unvoiced consonant *T*. In *bat* the *AE* is therefore shortened to 9 frames (146 ms).

The next rule checks for consonants occurring in clusters. If one consonant is followed immediately by another in the same word, the first is shortened to a subsegment in the same way that vowels are shortened by the previous rule.

The third rule adjusts the parameter trajectories between segments to make the transitions smoother and more natural. At each transition the frequency of each line spectrum parameter in the last frame of the leading segment and the first frame of the following segment are adjusted toward each other by an amount equal to one-third of the frequency distance between them. The trajectories are then smoothed for two frames on either side of the boundary.

The final rule adjusts the amplitude and voicing levels at the onset of utterance-initial vowels and at the end of utterance-final vowels to make the transitions to and from silence less abrupt and mechanical and thus generate more natural-sounding speech. If the first frame of an initial vowel is too strong, it can give the impression of an initial stop consonant. To eliminate this, the amplitude of the first two frames is reduced to produce a gentle onset, and the proportion of unvoiced excitation is increased in the first frame to give a slightly breathy quality to the start of the vowel. (See the Synthesis Algorithm section below for a discussion of the mixed excitation used in this system.) Likewise the ends of utterance-final vowels are softened by reducing the amplitude in the last two frames and increasing the unvoiced excitation in the final frame. In addition, the parameter trajectories in the final frame are adjusted to move toward a neutral position (i.e., a flat spectrum).

Timing Rules

In English, the duration of a given segment is affected by the stress level of the syllable in which it occurs and its position in the word and phrase. Segments in stressed syllables are longer than those in unstressed syllables; likewise, segments occurring at the ends of words or phrases are longer than those occurring in other positions. Segmental durations were studied in detail in the development of the MITalk synthesis-by-rule system [7]. Those timing rules have been adapted for use in this system.

In determining the duration of each segment in an utterance, the timing rules consider the stress level of the syllable, location of the segment in the word and phrase, and its phonemic class and context. (Stress levels for each syllable are assigned by the user after the sequence of segments has been specified.) Each rule prescribes a percentage increase or decrease in the duration of a segment in a particular context. The amount of compression or expansion varies depending on the factors listed above. Some sounds are more compressible than others, but no segment may be compressed more than the maximum amount specified in its header.

The timing rules for the MITalk system determine adjustments to the duration of each segment in milliseconds. For use in this system, the rules were modified to specify adjustments to the number of output samples generated in each frame. In this way the spectral pattern for a given sound is maintained when it is lengthened or shortened by the timing adjustment rules. For example, if the stored trajectory for a given LSP parameter rises by 400 Hz over the course of the sound, that same rise will occur regardless of the amount of compression or expansion. The slope of the trajectory will vary, but the change in frequency is constant.

Maintaining a constant frequency change regardless of the length of the sound could result in over-articulation, especially with tense vowels (*EE, EY, AY, OW, UW*) in unstressed syllables. To avoid this problem, utterances are entered phonetically rather than phonemically, using reduced vowels in unstressed syllables (e.g., *before* is entered as *B-EX-F-OR*, where *EX* represents the schwa sound /ə/, rather than as *Bf-EE-F-OR*).

This system can also vary the overall speaking rate of each utterance by changing the number of samples in the default synthesis frame. For a normal speaking rate, the default synthesis frame is 130 samples long. For slower speech, this can be increased up to 180 samples; for faster speech, the number of samples per frame can be reduced to 100. In the current implementation, the desired speaking rate is selected by the user. In the future it could be varied automatically depending on the type of material being presented.

Intonation Contours

To avoid an unpleasantly robotic monotone, particularly with longer phrases or sentences, it is important that synthesized speech have a smooth, natural intonation contour. Some synthesizers use an intonation model known as the "hat pattern" [11]. Basically, this model has three baselines that correspond to three different levels of stress (primary, secondary, and unstressed), with the pitch for each syllable following the appropriate baseline. Another approach that has been described models intonation patterns as a sequence of targets with various transitions depending on the relative heights of successive targets and their separation in time [12]. Transitions are determined by using quadratic equations and a set of constants and scaling factors.

In this system the intonation contour is determined by using the catenary equation that describes the curve of a chain or rope suspended between two points as illustrated in Fig. 3. This equation was chosen because it models several properties observed in natural intonation contours.

With the origin located a distance a below the vertex of the curve, the catenary equation is given as

$$y = a \cosh \frac{x}{a}. \quad (1)$$

The length of the arc from the vertex to any point on the curve is given by

$$s = a \sinh \frac{x}{a}. \quad (2)$$

In natural speech, pitch contours follow a sagging curve between successive stressed syllables, with the slope of the pitch change and the minimum pitch value dependent on the amount of time between the stresses [12]. Realistic synthetic pitch contours can be generated by using a proportional dip d so that as the separation between the end points of the curve increases, the difference between the highest and lowest points increases. Informal listening tests indicate that a fixed d equal to one-quarter of the horizontal distance between the endpoints results in satisfactory pitch contours. Let (X_1, Y_1) and (X_2, Y_2) represent the endpoints of the curve, and (X_0, Y_0) represent the origin. Given the proportional dip

$$d = \frac{(X_2 - X_1)}{4}, \quad (3)$$

the length of the arc is

$$L = X_2 - X_1 + d. \quad (4)$$

When the endpoints are of equal height this reduces to

$$L = 2s = 2a \sinh(x/a) \quad (5)$$

and

$$y = a + d. \quad (6)$$

The value of a can therefore be determined by

$$a = \frac{(L^2/4d) - d}{2}. \quad (7)$$

From these equations the location of the vertex can be computed as

$$X_c = \frac{1}{2} \left(X_1 + X_2 - 2a \operatorname{arcsinh} \frac{Y_1 - Y_2}{2a \sinh \frac{X_1 - X_2}{2a}} \right) \quad (8)$$

and

$$Y_o = Y_2 - a \cosh \left(\frac{X_2 - X_o}{a} \right). \quad (9)$$

The pitch at any point X along the curve is then

$$\text{pitch} = a \cosh(X/a) + Y_o. \quad (10)$$

When the end points of the catenary are not of equal height, the lowest point on the curve is not centered, instead it is closer to the lower end point. If the difference in height is sufficiently large, the lowest point on the curve will correspond to the lower end point. This is also consistent with natural pitch contours observed between stressed syllables of unequal level.

Two baselines are used to assign target values to each stressed syllable—the top line for stressed syllables with pitch accent, and the second, lower line, for stressed syllables without pitch accent. These target values are used as the Y coordinates of the end points of the catenary curve. (Although evidence exists to support the presence of low targets [12, 13], this system currently includes only high targets.) A third baseline is used as a floor, thus preventing the pitch from falling too low between widely spaced targets. These baselines are similar to those of the hat pattern model mentioned above, and they define the pitch range for the utterance. Each baseline exhibits a linear fall over the course of the phrase, with the top line falling more steeply than the bottom line, since it has been shown that for neutral declarative phrases, the pitch range narrows and drops over the course of the phrase [14].

The pitch contour is computed as a series of curves connecting successive stressed syllables. For each curve, the starting frames of the two target syllables are located. For the first three frames of the curve, the pitch follows the appropriate baseline; the curving pitch transition begins at the fourth frame of the first syllable [12]. The fourth frame of the first syllable and the first frame of the second syllable thus form the endpoints of the catenary curve. Boundary targets are inserted at the beginning and at the end of the utterance and at major phrase breaks to control the pitch at the onset and offset of each phrase. Another target is inserted at the end of the last accented syllable of the utterance (the nuclear stress, or tonic syllable) to control the posttonic pitch contour. The level of this special target and the shape of the contour are governed by the type of phrase: for

declarative phrases the pitch falls, so the target is low; for interrogative or continuation phrases the pitch rises, so the target is higher.

At major phrase boundaries within an utterance, targets are placed on the final frame of the segment preceding the boundary to generate the appropriate pitch rise, and they are also placed at the first frame of the following phrase. A seven-frame pause (114 ms at the normal speaking rate) is inserted, and the baselines at the beginning of the second phrase are reset [15]. To reset the baselines, the level of each line is raised by 50% of the difference between the current level and the level of that line at the beginning of the previous phrase. The end points of the baselines are constant, so that each phrase exhibits progressively less pitch declination.

Amplitude Contours

For each utterance, an overall amplitude contour is computed, which falls linearly over the course of the utterance. This contour is then shaped according to the stress levels of individual syllables. Syllables with primary stress receive full amplitude; those with secondary stress are reduced by 10%, and unstressed syllables are reduced by 20%.

The amplitude contour is further refined to reflect the inherent differences in amplitude between different types of sounds. The header of each segment contains an amplitude adjustment factor that indicates the ratio (in percent) of the amplitude of that segment to the utterance amplitude level at any given point. All vowels are 100%, meaning that the loudest frame in the vowel is set equal to the level of the sentential amplitude contour. Consonants are not as loud as vowels, so the adjustment factor for consonants ranges from 80% for liquids (*L, R, W, Y*) to 30% for voiced stops (*B, D, G*) and the softer voiced fricatives (*V* and *DH*; *Z* is inherently louder, and has an amplitude adjustment factor of 50%).

Synthesis Algorithm

After the timing, pitch, and amplitude adjustments have been made, the concatenated trajectories are output through the LSP synthesis algorithm. The algorithm in this system uses a mixed excitation and a proportional voicing indicator [16, 17] instead of the traditional binary voiced-unvoiced decision. The proportional voicing indicator specifies the amount of periodic (voiced) and aperiodic (unvoiced) excitation for each frame, and it is calculated from the locations of the first and second line spectrum frequencies. If a sound is voiced, a strong resonance is present in the lower spectrum (under 1000 Hz), so the first two frequencies are relatively low and close together. For unvoiced sounds, no low-frequency resonance is present, so the first two line spectrum frequencies are somewhat higher and more widely spaced.

SYSTEM PERFORMANCE EVALUATION

Two tests were conducted to measure the segmental intelligibility of the synthesized speech. The first was a diagnostic rhyme test (DRT), which uses a set of 224 single-syllable words to test the intelligibility of word-initial consonants [18]. The words are paired, with the members of each pair differing in only one attribute (e.g., *goat-coat* [voicing]; *nip-dip* [nasality]). A single DRT list was synthesized and scored by a panel of eight trained listeners. The overall score was 86.3% correct (corrected for guessing).

The second test was a modified rhyme test (MRT). This test uses 50 sets of 6 words each to measure the intelligibility of both initial and final consonants [19]. The members of each set differ in initial sound (e.g., *bill-hill-fill-will-kill-till*) or final sound (e.g., *mass-map-math-man-mad-mat*), but not both. Thirty naive subjects listened to the full set of 300 words, preceded by a training set of 50 words chosen at random from the test items. The overall score was 86.7% correct (84.1% when corrected for guessing). Table 2 compares the intelligibility test scores for this synthesizer with previously published scores for several other systems. The DRT scores for the Prose system were obtained with model 2020 v1.2; MRT scores were obtained with V3.0. For

the DRT, the Votrax chip was tested in a Namal Type and Talk system; for the MRT it was in a Votrax Type'n'Talk.

Table 2 — Comparison of Published DRT [20] and MRT [21]
Scores for Several Speech Synthesis Systems

Voice	DRT (% correct)	MRT (% correct)		
		Initial	Final	Overall
Natural speech	95.6	99.5	99.4	99.4
DEctalk Paul	87.5	98.4	95.1	96.7
DEctalk Betty	92.4	96.6	92.1	94.4
NRL LSP system	86.3	84.3	89.2	86.7
Prose	81.2	92.9	95.7	94.3
Infovox	83.6	90.0	85.0	87.4
Votrax SC01 chip	65.9	67.4	77.7	66.2

Figure 4 shows the percent of error accounted for on each test by each consonant phoneme. Part of the difference in performance between the DRT and the initial contrasts on the MRT is that the DRT offers only two choices; the MRT offers 6, and it is therefore a more difficult test. Also, the distribution of consonants differs on the two tests. For example, *D* occurs 22 times out of 192 items on the DRT (11.4%) and 7 times out of 300 on the MRT (2.3%); *H* occurs twice on the DRT (1.0%) and 12 times on the MRT (4.0%); *CH* occurs 10 times on the DRT (5.2%) but not at all on the MRT. The six consonants accounting for the highest percentage of error are *T*, *N*, *B*, *D*, *M*, and *TH*. Further work is needed on these sounds, but because they tend to be among the hardest to distinguish both in synthesizers and in natural speech [22], this does not indicate any problem specific to this synthesizer.

The relatively high error rate for *B* on the DRT is partly because *B* occurs so many times on the test (20 times out of 192 items, or 10.4%; on the MRT it makes up only 4.7% of the initial consonants). Although *B* accounted for 22.9% of the errors on the DRT, the segmental error rate was only 15%. Half of these errors were results of confusions with *V*, a particularly difficult contrast even in natural speech. On the MRT, *B* accounted for 12.2% of the errors on initial contrasts, with a segmental error rate of 20.3%.

The phoneme *SH* accounted for 17.1% of the errors on the DRT, with a segmental error rate of 37.5%. All of these errors were confusions with *CH*. On the MRT, the *SH-CH* contrast was not tested in initial position; *SH* was contrasted with *K*, *T*, *P*, *B*, *M*, *W*, *F*, *H*, *L*, and *R*, and was identified correctly 100% of the time (*SH* did not occur in final position). The *SH-CH* confusion is not usually troublesome, therefore this problem will be investigated. Further refinement of the *SH* segment may be required.

CONCLUSIONS

In this investigation we have shown that LSP parameters are well suited for use in a synthesis-by-concatenation system. The primary advantages of these parameters are that they allow fine adjustments to the speech spectrum, yet they require few rules. Also, because they are in the frequency domain, they are directly related to the spectral composition of the speech.

Percent of Error Accounted for by Each Consonant

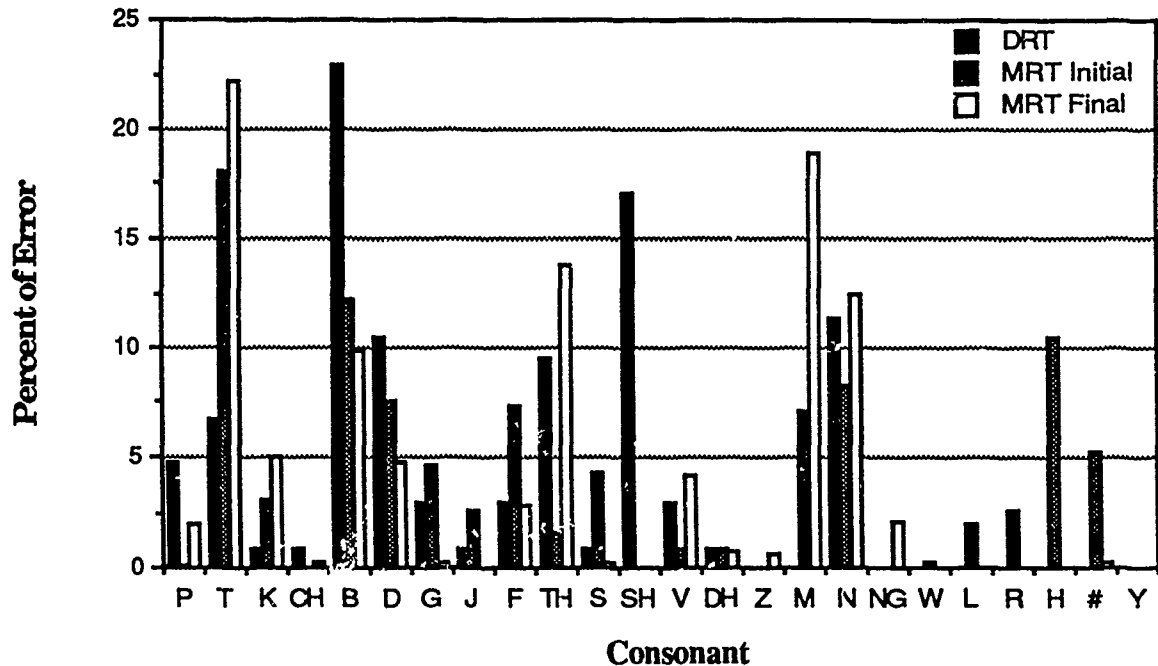


Fig. 4 — Percent of the total error on each test accounted for by each consonant phoneme. *DH* represents the voiced fricative sound at the beginning of *then*; # represents no consonant, such as at the beginning of *eel* or the end of *pay*.

This system has demonstrated good quality speech with a segment library created by using an 8 kHz sampling rate and 10 filter taps. These operating constraints were chosen to maintain compatibility with existing military voice communication systems. Some compensation for the limited bandwidth was made by reflecting the spectral components in the 2.5 to 4 kHz range upward into the 4 to 6.5 kHz range [16, 17]. This improves the speech quality somewhat, particularly for unvoiced consonants, but the use of a higher sampling rate and more filter taps would be expected to improve speech quality and intelligibility even further. The LSP synthesis algorithm would require minor alterations to accommodate these changes, and a new segment library would have to be collected, but the concatenation system itself, composed of the linguistic rules, the methods of adjusting the segmental timing and duration, and the determination of the pitch and amplitude contours, is essentially independent of the sampling rate and the number of filter taps used.

One of the primary applications of an unlimited vocabulary synthesis system is for text-to-speech conversion. Because the objective of this investigation was to demonstrate the feasibility of an LSP-based synthesis system, we chose to focus on the speech generation rather than the text analysis. A small text analysis package [23, 24] was integrated for demonstration purposes, but it was limited to single word utterances, and it did not generate stress levels. To make this system practical, a more powerful text analysis module will need to be integrated. Like any other text-to-speech system, this synthesizer would benefit from improved sentence parsing capabilities to provide more natural phrasing and stress placement. Improving the isochrony so the stressed syllables fall at approximately equal intervals would also help improve the quality of the speech, especially for longer phrases and sentences.

ACKNOWLEDGMENTS

This work was supported by NRL and the Office of Naval Research under Task Area RR021-05-42. The author thanks George Kang and David Tate for their advice and support throughout this investigation.

REFERENCES

1. D. H. Klatt, "Review of Text-to-Speech Conversion for English," *J. Acoust. Soc. Am.* **82**(3), 737-793 (1987).
2. F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Am.* **57** (S1) S35 (1975).
3. G. S. Kang and L. J. Fransen, "Experimentation With Synthesized Speech Generated From Line-Spectrum Pairs," *IEEE Trans. ASSP*, **ASSP-35**(4), 568-571 (1987).
4. G. S. Kang and L. J. Fransen, "Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)," NRL Report 8857, Jan. 1985.
5. J. B. Lovins, M. J. Macchi, and O. Fujimura, "A Demisyllable Inventory for Speech Synthesis," in *Speech Commun. Papers*, J. J. Wolf and D. H. Klatt, eds. (Acoust. Soc. Am., New York, 1979), pp. 519-522.
6. R. Schwartz, J. Klovstad, J. Makhoul, D. Klatt, and V. Zue, "Diphone Synthesis for Phonetic Vocoding," Proc. 1979 IEEE Intl. Conf. ASSP, 1979, pp. 891-894.
7. J. Allen, M. S. Hunnicutt and D. H. Klatt, *From Text to Speech: The MITalk System* (Cambridge Univ. Press, Cambridge, UK, 1987).
8. G. E. Peterson and I. Lehiste, "Duration of Syllable Nuclei in English," *J. Acoust. Soc. Am.* **32**(6), 693-703 (1960).
9. N. Umeda, "Vowel duration in American English," *J. Acoust. Soc. Am.* **58**(2), 434-445 (1975).
10. J. Olive, "A Scheme for Concatenating Units for Speech Synthesis," Proc. 1980 IEEE Intl. Conf. ASSP, 1980, pp. 568-571.
11. J. 't Hart and A. Cohen, "Intonation by Rule: A Perceptual Quest," *J. Phonetics* **1**, 309-327 (1973).
12. J. Pierrehumbert, "Synthesizing intonation," *J. Acoust. Soc. Am.* **70**(4), 985-995 (1981).
13. M. D. Anderson, J. B. Pierrehumbert, and M. Y. Liberman, "Synthesis by Rule of English Intonation Patterns," Proc. 1984 IEEE Intl. Conf. ASSP, 1984, pp. 2.8.1-2.8.4.
14. W. Cooper and J. Sorensen, "Fundamental Frequency Contours at Syntactic Boundaries," *J. Acoust. Soc. Am.* **62**, 683-692 (1977).
15. S. J. Eady, B. C. Dickson, S. C. Urbanczyk, J. A. W. Clayards and A. G. Wynrib, "Pitch Assignment Rules for Speech Synthesis by Word Concatenation," Proc. 1987 IEEE Intl. Conf. ASSP, 1987, pp. 1473-1476.

16. S. S. Everett, "Vocabulary Synthesis Based on Line Spectrum Pairs," NRL Report 9160, Jan. 1989.
17. G. S. Kang and S. S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 2: Synthesis Improvements," NRL Report 8799, June 1984.
18. W. D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Recognition*, M. E. Hawley, ed. (Dowden, Hutchinson and Ross, Stroudsburg, PA, 1977).
19. A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set," *J. Acoust. Soc. Am.* 37(1), 158-166 (1965).
20. R. L. Pratt, "Quantifying the Performance of Text-to-Speech Synthesizers," *Speech Tech.*, March/April 1987, pp. 54-64.
21. D. B. Pisoni, H. C. Nusbaum, and B. G. Greene, "Perception of Synthetic Speech Generated by Rule," *Proc. IEEE* 73(11), 1985, pp. 1665-1675.
22. J. S. Logan, D. B. Pisoni, and B. G. Greene, "Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems," *Research on Speech Perception Progress Report No. 11*, Indiana University (1985).
23. S. R. Hertz, J. Kadin, and K. J. Karplus, "The Delta Rule Development System for Speech Synthesis from Text," *Proc. IEEE* 73(11), 1985, pp. 1589-1601.
24. Hertz, S., "English Text to Speech Conversion with Delta," *Proc. 1986 IEEE Intl. Conf. ASSP*, 1986, pp. 2427-2430.